Cache Strategies For High-end PC and Servers

Name: David Bondurant
Title: Vice President of Marketing
Company: Enhanced Memory Systems

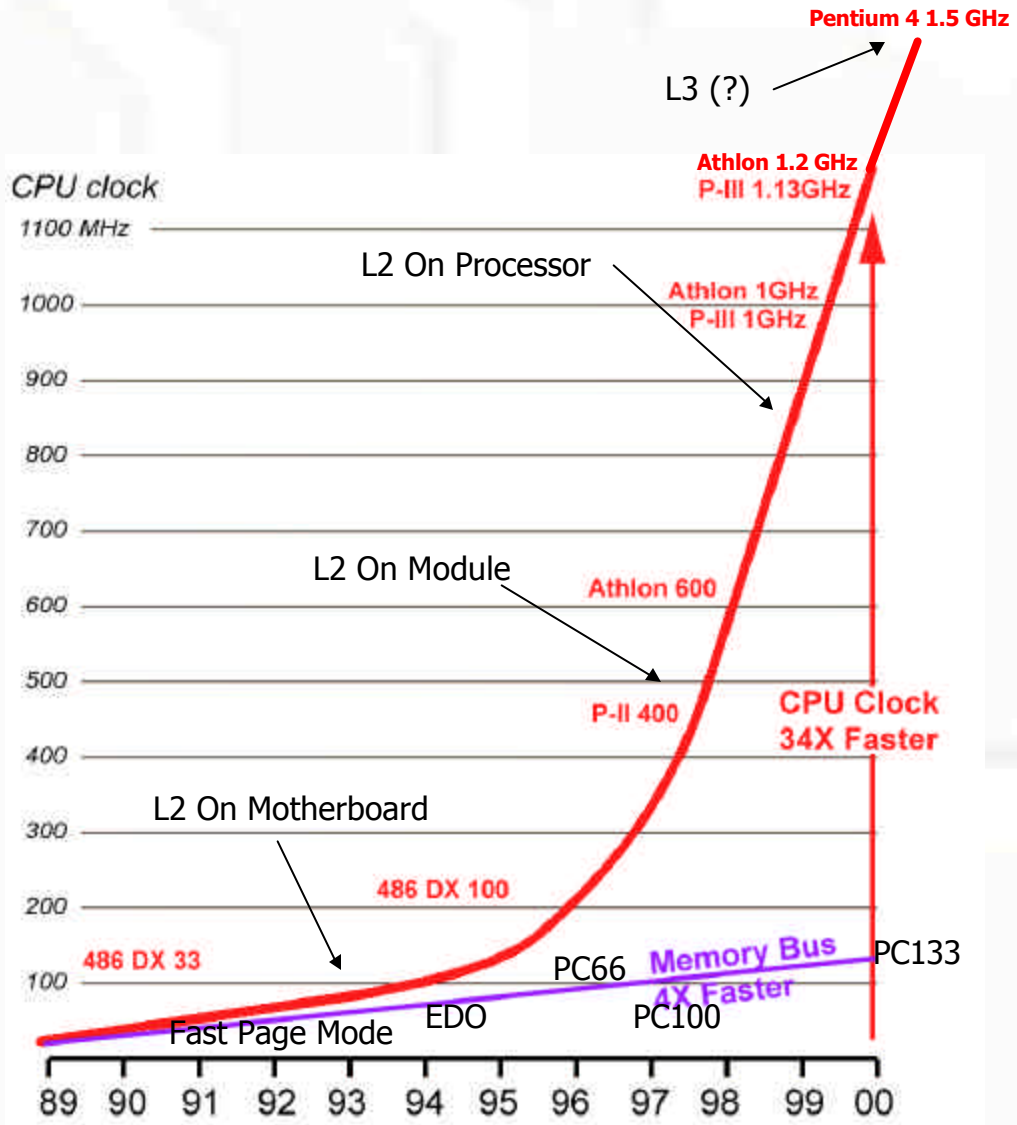
San Jose January 23-24, 2001 • Taipei February 14-15, 2001

Presentation Outline

- Cache Background
- Why Caches Don’t Work
- Proposed Cache Alternatives
- L3 Cache In Main Memory
  - Enhanced DDR

Evolution of Caching In Computer Systems

- DRAM Bandwidth Keeps Pace With Processor Speed
- DRAM Latency Became More than 5x Worse
- Multi-Tier Cache Hierarchy Develops



Platform Conference

Enhanced Memory Systems Inc.

What is a Cache?

- Fast Memory That Holds A Subset of Main Memory For Faster Access on Subsequent References To The Same Location

Why Do Caches Fail?

- Cache Not Large Enough To Hold Currently Executing Programs and Data (Memory Footprint > Cache Size)
  - Constant Cache Misses, Increased Memory Bandwidth Requirement
- Frequent Program or Applications Changes (Multi-Tasking, Real-time Operation)
  - Cache Thrashing
- Dynamic Datatypes (Graphics, Streaming Audio/Video)
  - Non-Cachable Data

Growing Memory Footprint

- Larger Operating Systems
  - 100KB to >60MB
- Larger Applications Programs
  - 10KB ? 5 MB (Word/Excel) to 15 MB (Photoshop)
- Larger Databases
  - 1KB Text File ? 100MB Graphics Files ? 1GB Video Files

Real-Time Operating Environment

- Preemptive Multi-Tasking Operating Systems
- Numerous System Extensions Executing In Background
- Concurrent Audio/Video Streaming

Changing Applications and Datatypes

- Productivity Age (1986-1994) ? Word Processing, Spreadsheets
- Internet Age (1995-2000) ? Web Browsers, Email
- Digital Lifestyle Age (2001-) ? Digital Imaging & Video, Streaming Audio, Streaming Video

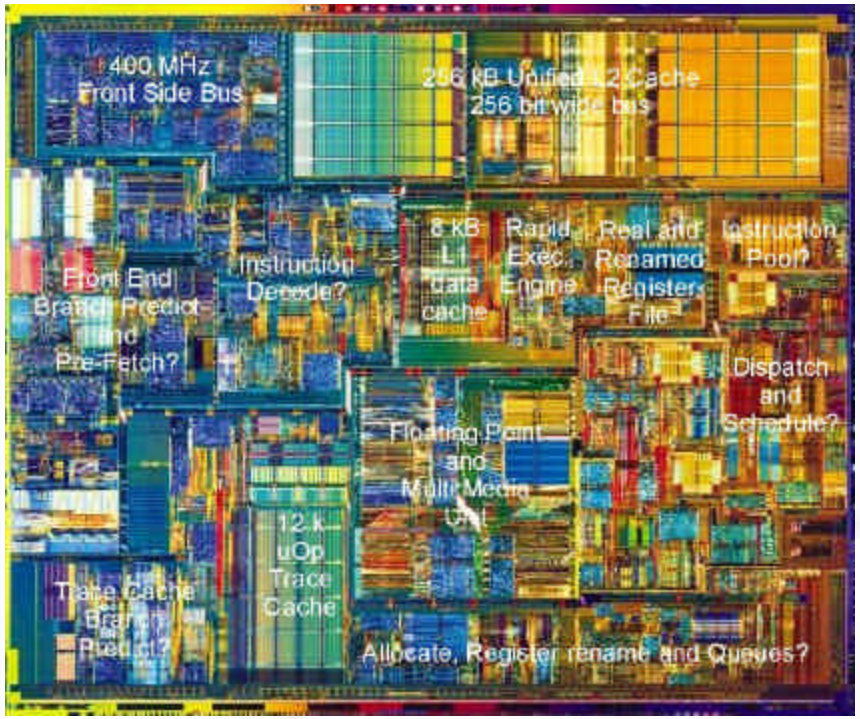
Source: Steve Jobs, MacWorld 2001

Cache Options

- Larger L2 On-Chip
- L3 Cache On Backside Bus
  - IBM Power4
- L3 Cache On Chipset
  - Micron's Proposed Chipset
- Cache In Main Memory
  - Enhanced Memory Systems EDDR

Larger L2 On-Chip

- Production Cost and Power Constraints Limit Integrated Cache To 256KB Today
- Larger On-Chip Cache Expected In Next Generation (1-2MB)



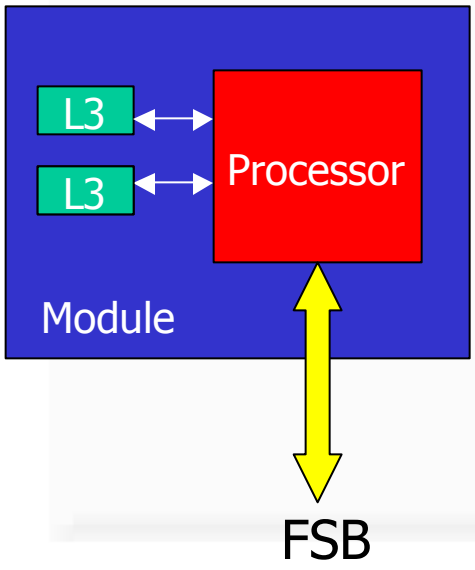
Pentium 4 Die

Platform Conference

Enhanced Memory Systems Inc.

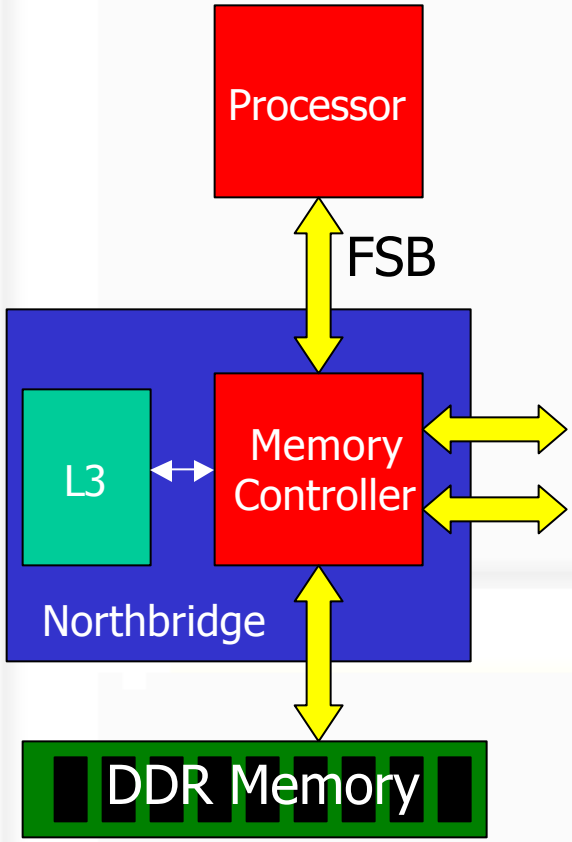
L3 Cache On Backside Bus

- First Used In High-end Server Systems
  - IBM Power4
- Benefits
  - Higher Bus Speed, Lower Latency
  - No Bus Contention
- Disadvantages
  - Expensive Multi-Chip Packaging
  - Expensive DDR SRAM Components



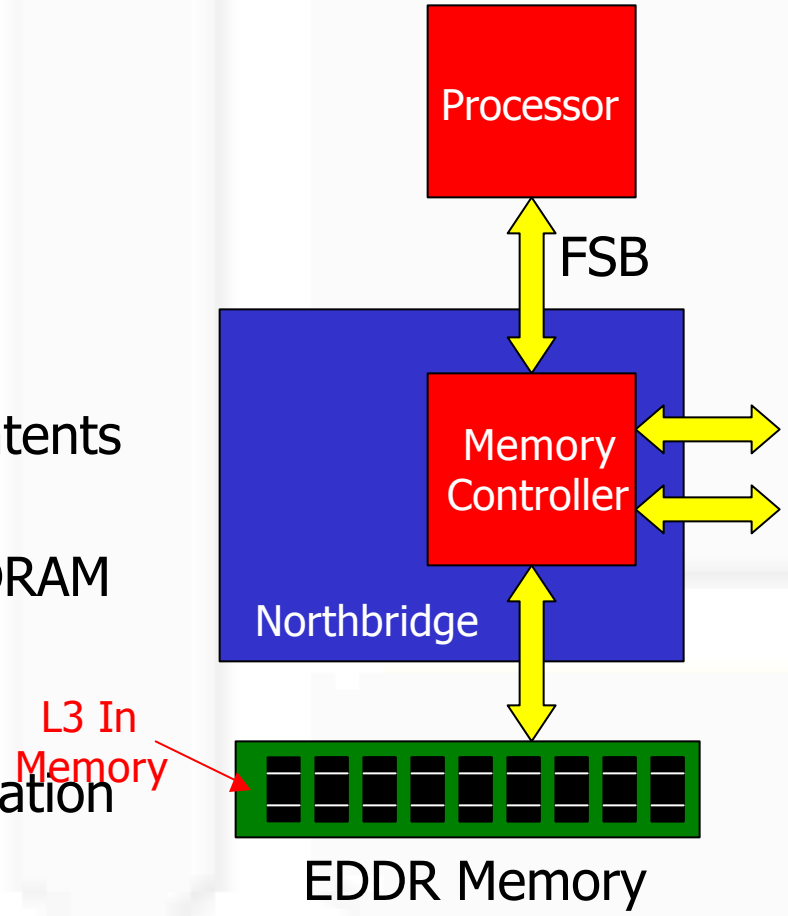
L3 Cache On Chipset

- Proposed For Future Pentium 4 Chipsets
  - 2 to 8MB Embedded SRAM
- Benefits
  - Simple Packaging
- Disadvantages
  - FSB Latency and Bandwidth Limitations
  - Costly Embedded SRAM Could Double Chipset Cost



L3 In Main Memory

- Enhanced DDR (EDDR)
  - Current JEDEC DDR Superset
- Benefits
  - Little Additional Cost To System
  - Improves Latency and Effective Bandwidth To Entire Memory Contents Not A Limited Subset of Data
  - Upward Compatible With DDR SDRAM Module Pin-out and Functions
- Disadvantage
  - FSB Latency and Bandwidth Limitation

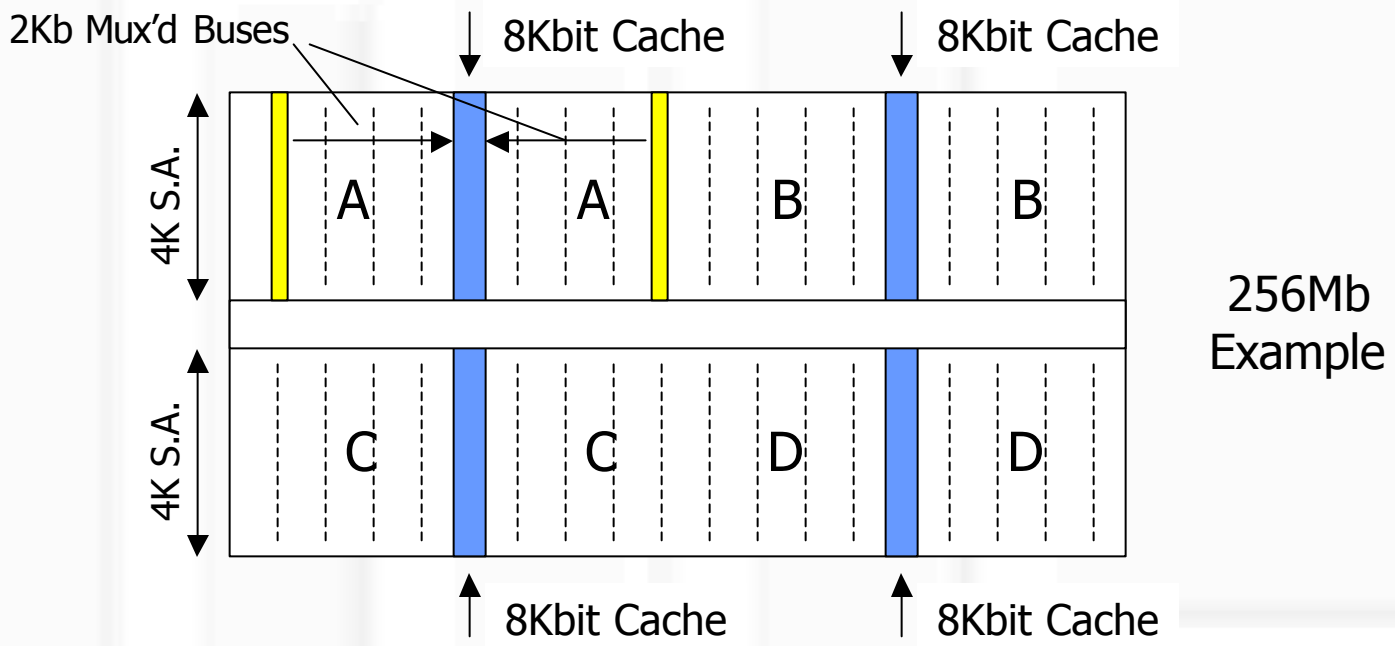


Cache Economics 101

- DRAM is 10x More Cost Effective Than SRAM
- DRAM is 100-200x More Cost Effective Than Processor or Chipsets
- Where Should Cache Be Integrated For Best Price/Performance?

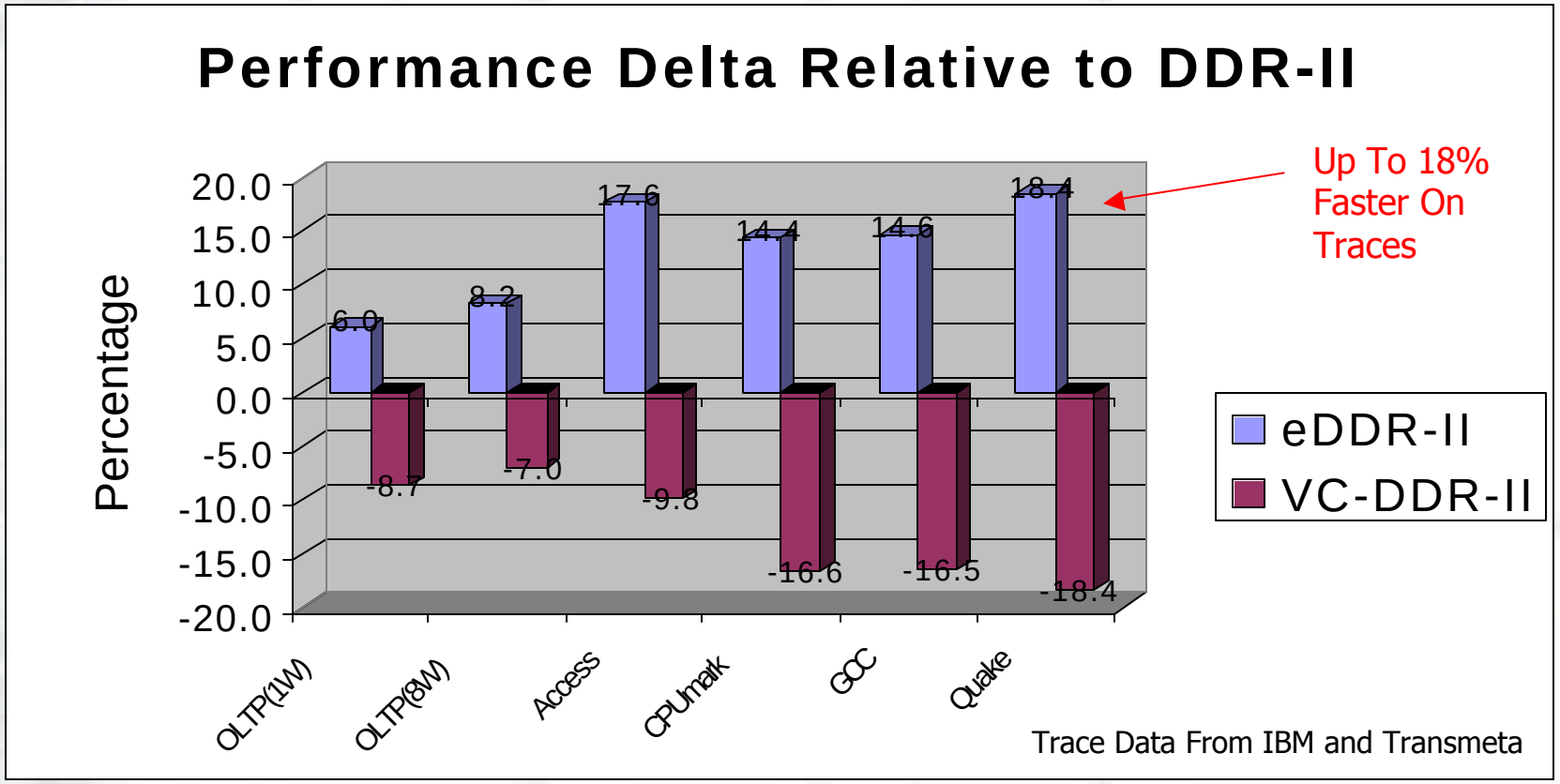
	Transistor Count	Price	Cost/Transistor
Chipset	1.0E+06	$20.00	20.00
P4 Processor	4.8E+07	$750.00	15.63
P3 Processor	2.4E+07	$150.00	6.25
8 Mb DDR SRAM	4.8E+07	$50.00	1.04
128MB SDRAM DIMM	1.0E+09	$60.00	0.06
			Microcent/Transistor

EDDR Architecture: Cache In DRAM



- **Central 8Kbit Row Caches (One Per Bank) For Low Chip Overhead (1-3%)**
- **Compatible With DDR Functions, Timing, Pin-outs**
- **Up To 22% Performance Advantage**

University of Michigan/Maryland JEDEC Future DRAM Study Results

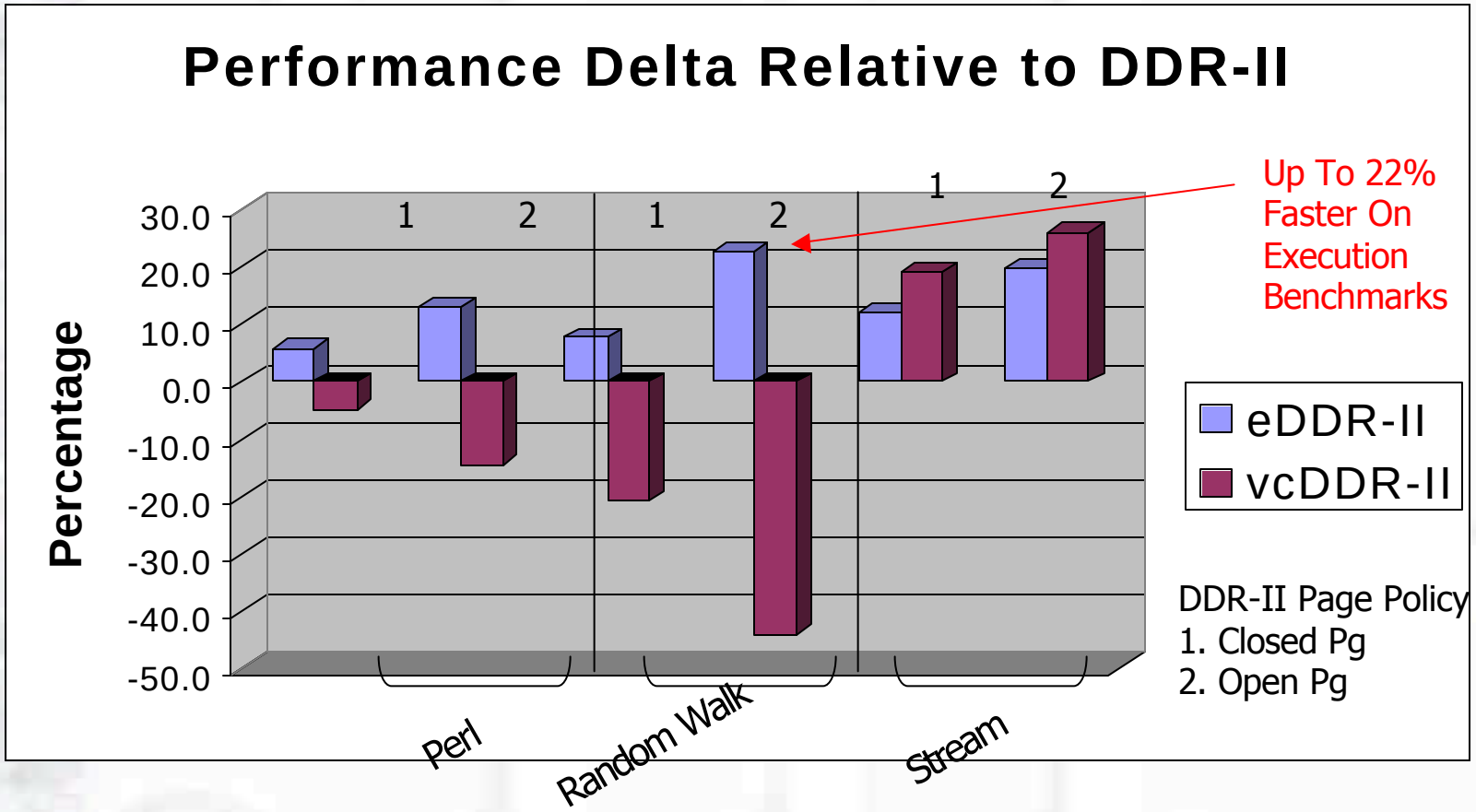


http://www.eecs.umich.edu/~btdavis/papers/mem_wall.isca2k.pdf

Platform Conference

Enhanced Memory Systems Inc.

University of Michigan/Maryland JEDEC Future DRAM Study Results



http://www.eecs.umich.edu/~btdavis/papers/mem_wall.isca2k.pdf

Platform Conference

Enhanced Memory Systems Inc.

What Do The Experts Say?

Maurice V. Wilkes
AT&T Research Labs, Cambridge, UK

The Memory Gap
Solving The Memory Wall Problem Workshop
Keynote Address
Vancouver, BC June 11, 2000

"Memory Latency leads to a significant and increasing amount of processor idle time. ……. If the memory latency remains unchanged, the number of cycles of processor idle time is doubled with each doubling of the speed of the processor.

The ideal solution would be the development of a new memory technology that would lead to memories with much lower inherent latency than present ones, and thus deal with the problem at its source."

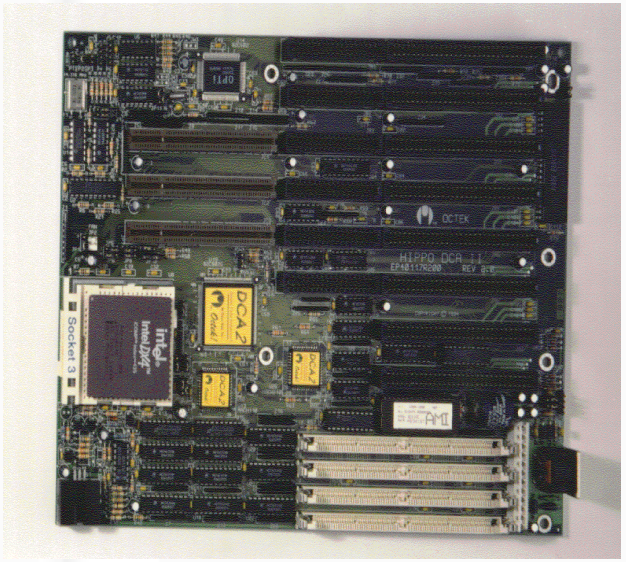
http://www.ece.neu.edu/conf/wall2k/wilkes1.pdf

Platform Conference

Enhanced Memory Systems Inc.

déjà vu – Softwin Reports (Dec 1994)

- "Notwithstanding the considerable advantage of the Pentium architecture, the DCA/2 - and its 16 Megabyte complement of 15/35 ns EDRAM - soundly beat Intel's Pentium system board at both 16 and 32 bit operations. The DCA/2 is the only motherboard to ever meet Softwin Laboratories 32 bit qualification standard."



  - Comparison of Ocean DCA/2 486DX4/100 With EDRAM Main Memory To Pentium 100 Motherboard With L2 Cache and DRAM

Summary

- Current Microprocessors Integrate L1 and L2 Caches To Offset Slow Latency of Main Memory
- Microprocessor Performance Still Limited By External Memory Latency
  - Growing Memory Footprint
  - Multitasking
  - New Datatypes (Graphics, Audio, Video)
- External L3 Caching Will Be Required In High-end PC and Server Systems
- L3 Cache In DRAM Is Shown To Offer The Best Price/Performance At The System Level
- Enhanced DDR (EDDR) Architecture Is Already a JEDEC DDR Superset